Text Mining

Task 1 & 2

Xiang Li

12/3/2021

Task 1- Pick A Book

I choose `My Doggie and I` by R.M.Ballantyne as the resource of my text analysis. `My Doggie And I` tells the story of John Mellon (almost a doctor in chapter one) and what happens to him after he meets a certain little canine. This story surrounds a child waif, a young woman, a young gentleman doctor, and an elderly lady. This tale unfolds the story of a bond that brings these unlikely friends together and merges their separate paths of life into one common path.

First, let us get a basic sense of most common word in the book.

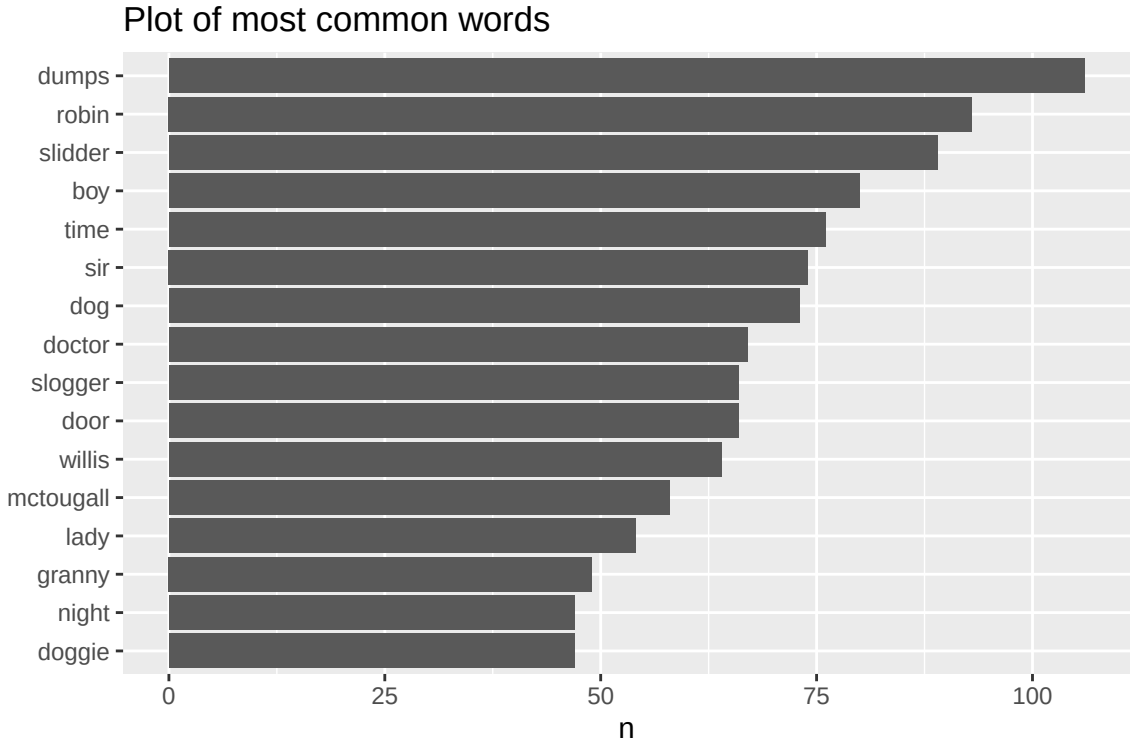


Figure 1: A visualization of the most common words

Task 2- Bag of Words Analysis

With several options for sentiment lexicons, we can use inner_join() to calculate the sentiment in different ways. Following the step in `Textmining in R`, I choose three general-purpose lexicons are * `AFINN` from Finn Årup Nielsen, * `bing` from Bing Liu and collaborators, and * `nrc` from Saif Mohammad and Peter Turney. Then, I generate three plot which showed an estimate of the net sentiment (positive - negative) in each chunk of the novel text for each sentiment lexicon.

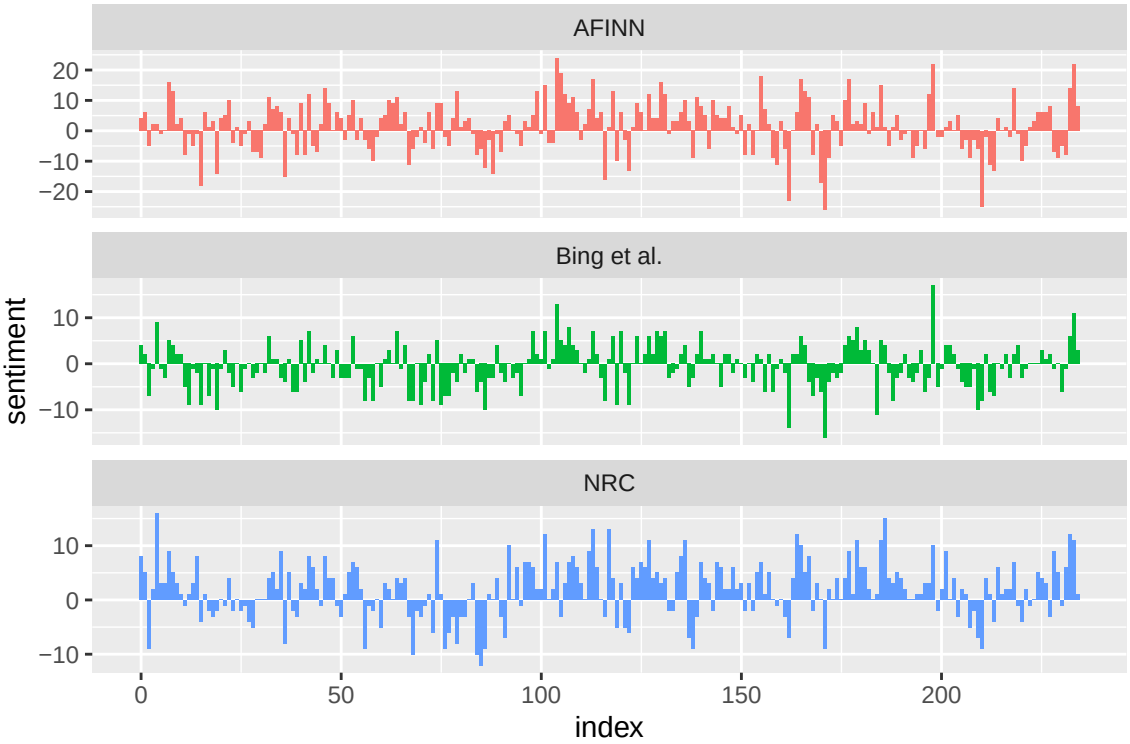


Figure 2: visualization of an estimate of the net sentiment (positive - negative) for each sentiment lexicon

Discussion: I use integer division to define larger sections of text every 20 line in the book. Then I have three distinct lexicons for calculating sentiment flows, which produce findings that differ in absolute terms but follow similar relative paths across the novel. In the novel, I detect comparable drops and peaks in emotion at roughly the same locations, but the absolute numbers are much different. The `AFINN` lexicon has the highest absolute values and the highest positive values. `Bing et al` lexicon has lower absolute values and appears to mark longer chunks of continuous positive or negative text. The `NRC` findings are skewed upward in comparison to the other two, positively labeling the text, but identifies identical relative changes in the text.

One advantage of having the data frame with both sentiment and word is that we can analyze word counts that contribute to each sentiment. By implementing count() here with arguments of both word and sentiment, we find out how much each word contributed to each sentiment. Now let's look at the analysis for these bing and nrc lexicons.

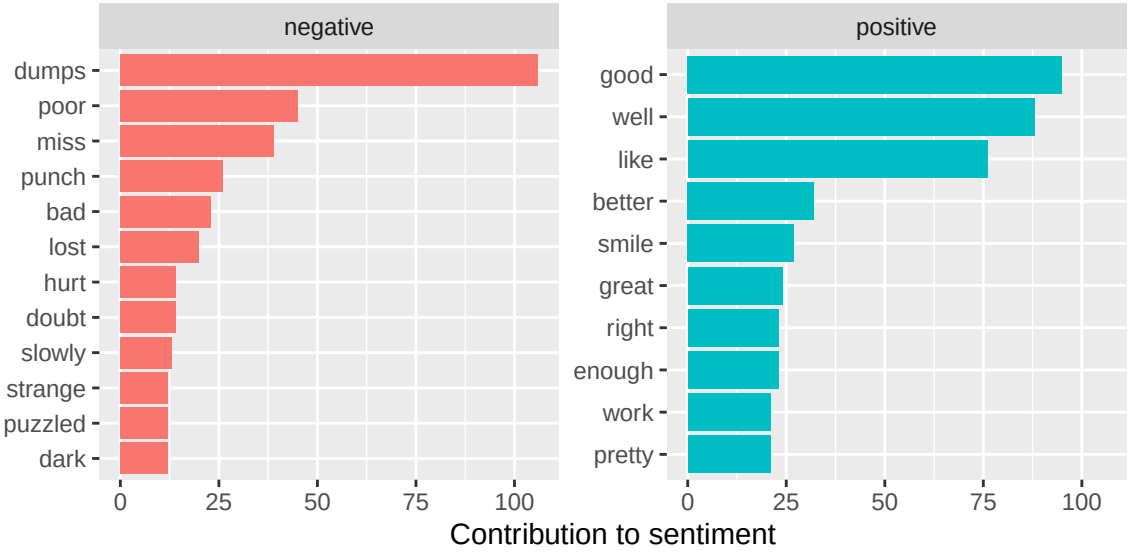


Figure 3: Most common positive and negative words in Bing Lexcions

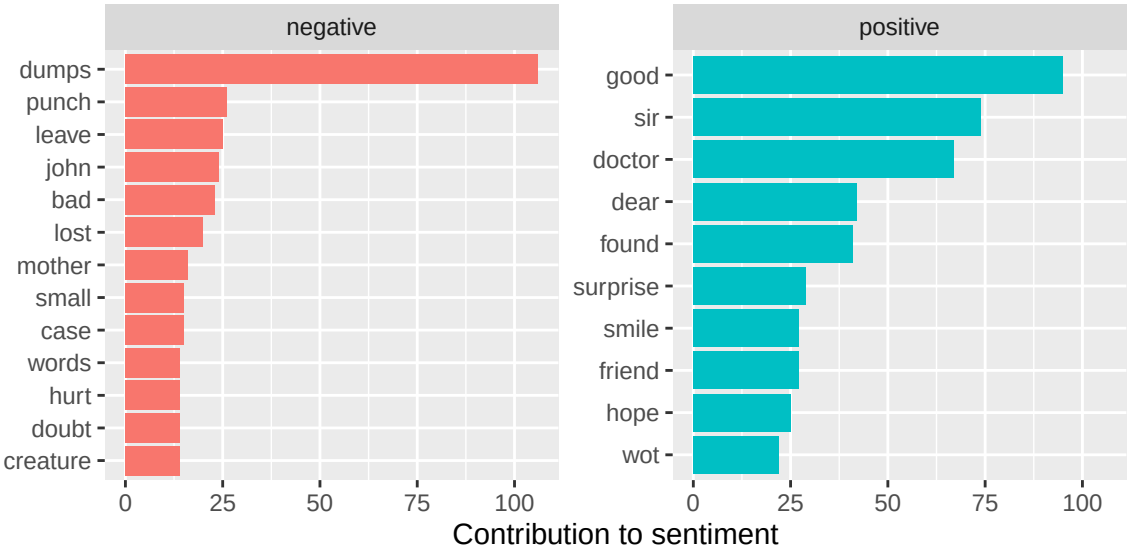


Figure 4: Most common positive and negative words in nrc Lexcions

Discussion: These two figures showed the most common negative and positive words in different lexicons. Different words have different contribution to sentiment in different lexicon. However, they both have `dumps` as most common negative words and `good` as most common positive words.

World Clouds

Let's see the worldclouds for the `My Doggie and I` using worldcloud package.

Figure 5: Worldclouds with most common words.

In this figure, it displays the most common words used in the book, which are `doggie`, `willis`,`slidder` and so on. This cloud matched the plotline in my book since they are main character and main scene in my book.

Let's see the worldclouds of most common negative and positive words for the `My Doggie and I`.

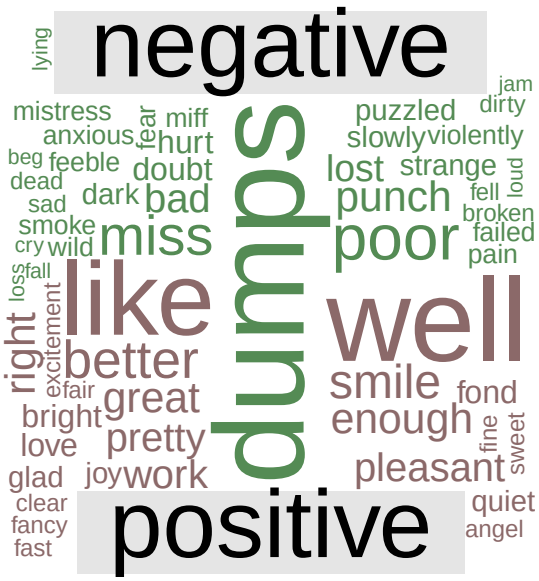


Figure 6: Worldclouds with most common negative and positive words.

In this figure, it displays the most common negative and positive words used in the book, which are `dumps`, `good`,`well` and so on. This cloud matched the plot line in my book since in the book, there are lots of dialogues and these words are often happened in the oral speaking.

Task 2 Extra credit

The additional lexicons I found is that `loughran`. And here is the sentiment flow of the book.

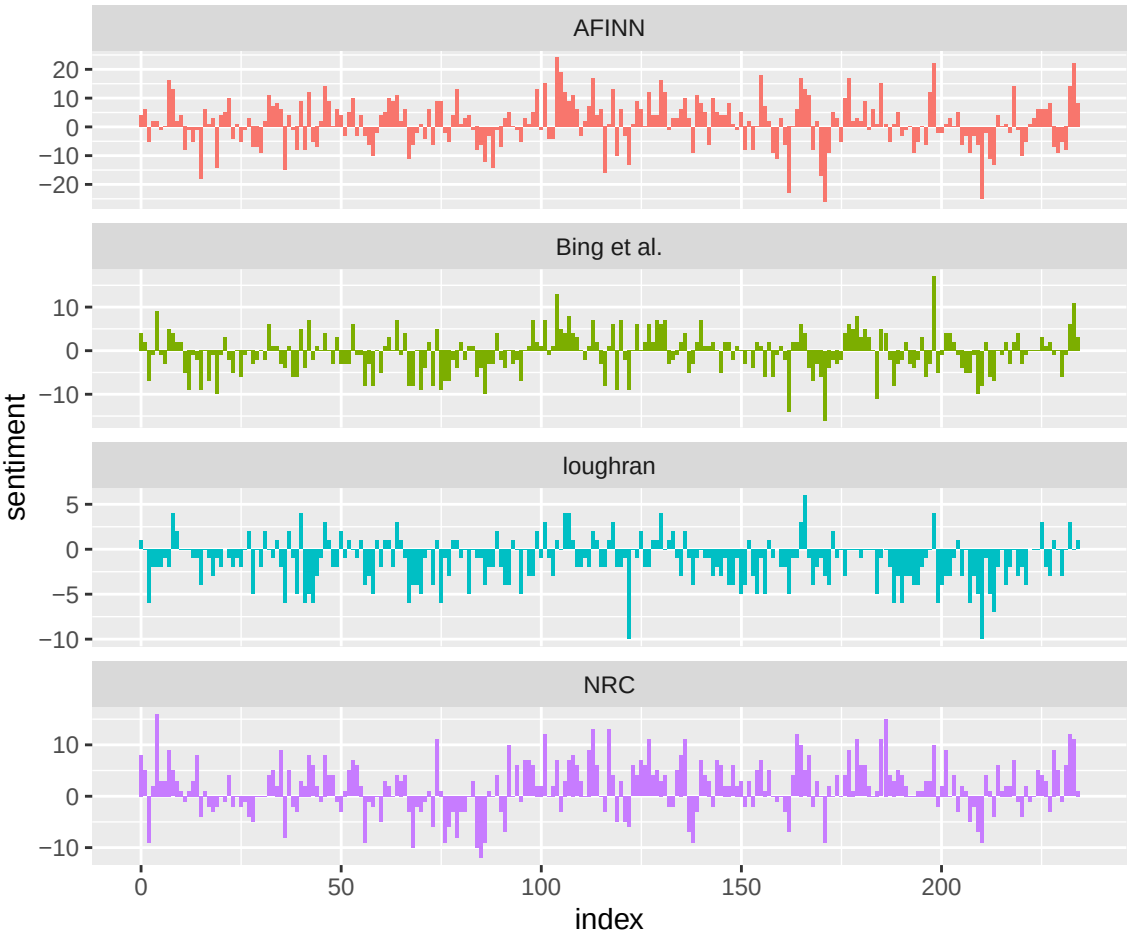
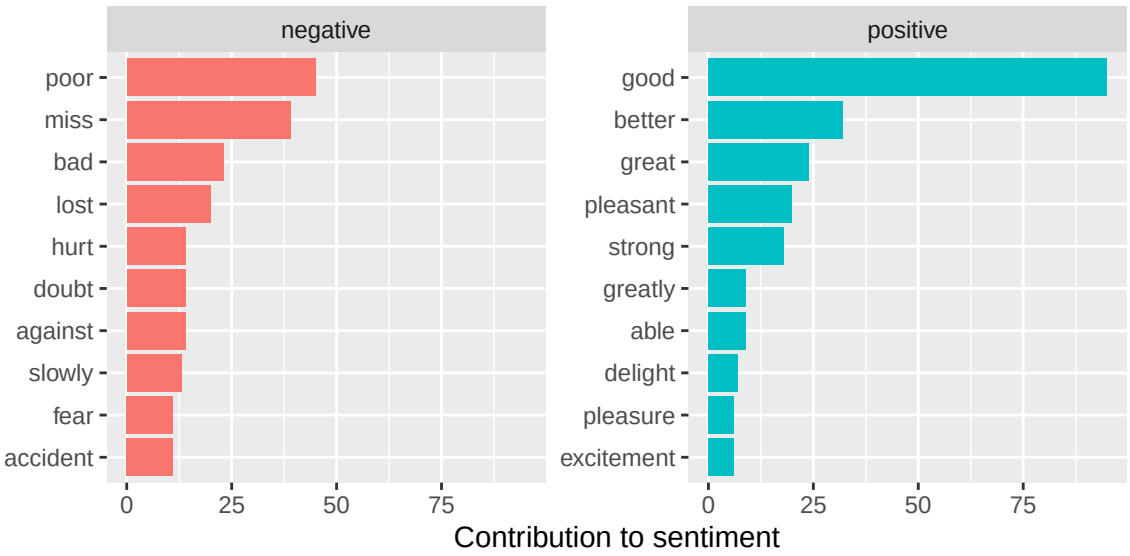


Figure 7: Using loughran to create a sentiment flow

Discussion:

Compared to previous three lexicons' result, the sentiment flow by `loughran` has the lower absolute values and the more negative values. Not surprisingly, this comparison plot adding `loughran` Lexicon produces findings that differ in absolute terms but follow similar relative paths across the novel. In the novel, I detect comparable drops and peaks in emotion at roughly the same locations, but the absolute numbers are much different.

Here is the plot which showed how the most positive and negative words contributing to sentiment. Not like the previous two lexicons, the most negative word in `loughran` Lexicon is `poor`.


negative
poor
miss
bad
lost
hurt
doubt
against
slowly
fear
accident
0
25
50
75
positive
good
better
great
pleasant
strong
greatly
able
delight
pleasure
excitement
0
25
50
75
Contribution to sentiment